*The work We submitted as part of this assignment is original, and due credit is given to others where appropriate. We accept and acknowledge that We are* ***solely*** *responsible for if the assignment found to be plagiarized in any way, and we will be subject to School’s Academic Integrity policy*

Group 10: Jiadai Yu(U63347232), Yuesen Zhang(U25148433), Xin Su(U07869307)

Date: 29th Aug 2022

Based on the correlation matrix, our group would suggest to include the price, year, mileage, engine size, and tax. The variables we did not include are mpg, transmission, and fuel type. Since the price is the factor the clients are most concerned about, our group focused on analyzing the regression analysis between the price and other variables.

From the matrix, we separated these variables into two types one is time-related, and the other is environment related. The time-related type has the variables year and mileage, and both of these two variables have a strong correlation with the price. Hence, these two variables, year and mileage, are chosen in the regression analysis. In addition, the environment-related type has the variables engine size, tax, mpg, transmission, and fuel type. Among these variables, the fuel type and transmission have the weakest correlation with the price, so we excluded them from the regression analysis. Moreover, both engine size and tax have a 0.41 correlation with the price, while the mpg has a -0.5 correlation with the tax. In other words, these three variables are dependent on each other. Therefore, we picked two of these three variables, engine size and tax, for our regression analysis.

Overall, our group suggests to include price, year, mileage, engine size, and tax in the regression analysis.

In this case, we keep all the explanatory variables the same, the only difference between Model 1 and Model 2 is the response variable. After linear regression and modeling, we obtain the models:

Model 1

price = 1.21292845e+03\*year - 6.04409697e-02\*mileage

+ 5.67107198e+03\*engineSize + 6.50246287e+00\*tax

- 2441011.0945755113

Model 2

log\_price = 1.30621388e-01\*year - 4.95493934e-06\*mileage

+ 3.92918890e-01\*engineSize + 2.47154619e-04\*tax

- 254.5480523731459

We know that in Python, R^2 represents the fraction of variance of the actual value of the response variable captured by the regression model, which is bounded between 0 and 1. If the value of R^2 is greater, it means the response variable captured by the model is better-observed. In other words, if the value of R^2 is higher, the model fits better with the dataset. For Model 1, we get the value of R^2 is 0.7045. For Model2, we get the value of R^2 is 0.7856. According to the definition and the comparison of the data, we can conclude that Model2 is more well-observed.

In addition, we also obtain the Mean Squared Error(MSE). We know that by definition the MSE represents the error of the estimator or predictive model created based on the given set of observations in the sample. Thus it is used to measure the quality of the model based on the predictions made on the entire training dataset vis-a-vis the true label/output value. If the value of MSE is higher then the model has more errors. If the value of MSE is lower then the model has less error and more accuracy. For Model 1, we get the value of MSE is 6715130.61. For Model 2, we get the value of MSE is 0.03535. We observed that the MSE of Model 2 is significantly smaller than the value of Model 1, and very close to 0. Which indicates the outstanding accuracy of Model 2.

Lastly, we can also compare the actual price of the test data and the predicted output. We know that the predicted value should be close to the actual value, then we can conclude the model is the best fit for the dataset. For Model 1, We observed that the actual price1 is $12495 and the predicted value is $15337, which has a 22.75% of the difference. For Model 2, we observed that the actual price2 is $12495 and the predicted value is $ 14238.18, which has a 13.95% difference. Then we can see that Model 2 has less percentage difference, and we can conclude that Model 2 is the best fit.

We would propose the second model to our client as the final recommendation, which is the linear regression model with production year, number of miles the vehicle traveled, the car's engine size, and annual tax as explanatory variables, and the logarithm of price as the dependent variable. The reason is not only that the statistical results of this model are better than first model, it also has a stronger probabilistic theoretical support.

From the scatter plot of production year and mile age versus price, and the prediction of test data, we can tell that the distribution of the dependent variable price is not linear. The reason why the first regression model achieved a seemingly satisfactory result that R^2 = 0.7, is due to the relatively small range of the data. Within this small distance, the relationship between price and the independent variables is roughly linear. If the client would like to promote the model to predict a larger range of data, the predictions are likely to be inaccurate.

The logarithm of price, on the other hand, shows a significant linearity in the scatter plots with both the production year and mile age of each vehicle. In addition, the test data results yielded by the second model predictions are evenly distributed on both sides of the linear regression. A linear regression model such as the second one is suitable for generalization to a larger range of data and is also more likely to produce valid pridictions in data sets beyond those used in the current study.

For further discussion, we can justify the log-price model with probability theory. In terms of the random data distribution, year and mile age are variables that describe " the time until an event occurs " which fit the exponential distribution. The independent variables, year and mile age, have an exponential level influence on price. After logging both sides of the exponential distribution equation, there is a linear correlation between log price and independent variable.

In summary, the log-price model provides more reliable and valid statistical results, is more suitable for generalization to a larger range of data, and adheres better to the theoretical knowledge of probability theory. We would recommend this model to my clients.